People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

Mohamed Khider University of Biskra

Faculty of Exact Sciences

Department of Mathematics

Thesis Submitted in Partial Execution of the Requirements of the Degree of

Master in "Applied Mathematics"

Option : Probability

By Dalal SAIDI

Title :

Dynamic Modeling of Sparse Longitudinal Data Using Stochastic Differential Equations

---

Examination Committee Members :

Mr.	Boubakeur LABED	Assoc.Prof.	U. Biskra	President
Mr.	Nabil KHELFALLAH	Professor	U. Biskra	Supervisor
Mrs.	Hanane BEN GHERBAL	Assoc.Prof.	U. Biskra	Examiner

03/06/2025

Dedication

To those who have been my greatest support after God in every step of my academic journey. **My dear parents, Mohamed said and Zoubida**, I offer you my deepest gratitude and appreciation for your unconditional support. You have always been my refuge and source of strength.

To **my beloved siblings, Hamida, Walid, Hicham, Riyadh, Saliha, Douâa, and Aymen,** you are the heartbeat that fuels my determination and perseverance. Your love and support mean the world to me, and I am forever grateful to have you by my side.

To **my esteemed professors**, who guided me with their knowledge and wisdom, illuminating my path and inspiring me to grow.

To everyone who has played a part in my success, I dedicate this achievement as a token of my love and gratitude.

Dalal SAIDI

Acknowledgment

First and foremost, I extend my deepest gratitude to **God** Almighty for His endless grace and blessings. He has paved the way for me, granting me the strength and wisdom to complete this work. All praise and glory be to Him, the Most High.

I would like to express my sincere appreciation and profound gratitude to my esteemed supervisor, **Nabil KHELFALLAH**, for his unwavering guidance, invaluable advice, and continuous support throughout this journey.

My heartfelt thanks also go to the distinguished members of the Examination Committee, **Boubakeur LABED** and **Hanane BEN GHERBAL**, for dedicating their time and effort to reviewing and evaluating my work with great care.

I am also deeply grateful to all my professors and teachers who have shaped my academic path, especially those in the mathematics department, whose knowledge and mentorship have been instrumental in my growth. Lastly, I want to acknowledge myself for believing in my abilities, for persisting through challenges, and for always striving to give more than I receive.

Notations and symbols

These are the different symbols and abbreviations used in this thesis.

Symbols:

$(\Omega, \mathcal{F}, \mathbb{P})$	:	Probability space.
$(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$	:	A complete filtered probability space.
T	:	A strictly positive number.
$X = (X_t)_{t\in[0,T]}$	:	A stochastic process.
$\mathcal{N}(\mu, \sigma^2)$	:	The normal distribution is defined by a mean μ and a variance σ^2.
$\mathbb{L}^2(0, T)$	:	Space of square-integrable functions on the interval $[0, T]$.
B_t	:	Brownian motion.
$\mathcal{M}^{n\times m}$	:	The space of all real matrices of dimensions $n \times m$.
$\mathbb{E}[.]$	:	Expectation.
$var(.)$	:	Variance.
$cov(x, y)$	:	The covariance between two random variables x and y.
$\mathbb{E}(. \mid .)$	:	Conditional mean.
$\|.\|$	:	The Euclidean norm in $\mathbb{R}^n$.
$var(. \mid .)$	:	Conditional variance.
$\mathbb{R}^d$	:	The $d-$dimensional Euclidean space, representing the set of all real vectors with d components.

$B^+(t)$	:	Reflected Brownian motion is a stochastic process defined as the absolute value of standard Brownian motion $B(t)$ such that $B^+(t) = \|B(t)\|$.
$\mathcal{F}_t^{X_0} \vee \sigma\{B_s : 0 \leq s \leq t\}$	:	The smallest σ-algebra generated by the initial condition X_0, and the Brownian motion $B_s$up to time t.
$\left\{\mathcal{L}(X_K), \mathcal{L}(\hat{X}_K)\right\}$	:	The distributions of X_K and its corresponding estimator $\hat{X}_K$ as $\mathcal{L}(X_K)$ and $\mathcal{L}(\hat{X}_K)$ respectively .
$(\mathcal{F}_t^X)_{t\geq 0}$	:	Natural filtration of the stochastic process X defined by $\mathcal{F}_t = \sigma(X(s), 0 \leq s \leq t)$.
$C^2([0,\infty) \times \mathbb{R})$	:	The space of functions that are twice continuously differentiable on the domain $[0,\infty) \times \mathbb{R}$.

Notations:

SDEs	:	Stochastic differential equations.
a.s.	:	Almost sure.
$\mathbb{P}-$a.s.	:	Almost sure in probability $\mathbb{P}$.
$m(\cdot)$	:	The mean function.
$v^2(\cdot)$	:	The conditional variance function.
i.e.	:	That's to say.
w.r.t.	:	With respect to.
a.e.	:	Almost every where.
DM	:	The proposed dynamic modeling approach.
LW	:	The covariance completion method.
$RMSE$	:	The root-mean-square error.
FPCA	:	Functional principal component analysis.

Contents

List of Figures

List of Tables

Introduction

In various scientific fields, mathematical models rely on longitudinal data to monitor temporal developments of different phenomena. However, sparse longitudinal data pose a significant challenge due to irregular observation times and missing information in certain periods. One effective approach to addressing these challenges is the use of stochastic differential equations (SDEs), which provide a dynamic framework for analyzing sparse data and making predictions based on limited observations. This approach serves as an alternative to traditional methods such as functional principal component analysis (FPCA), which require strong assumptions that may not be feasible in cases of sparse data.

The roots of functional data analysis date back several decades, with techniques such as FPCA and functional regression models being developed to study temporal processes. However, these methods face difficulties when data are incomplete or sparse. In recent years, stochastic differential equations have gained increasing attention as a tool for modeling dynamic stochastic phenomena. Notable contributions to FPCA have been made by Kleffe (1973)[12], Castro et al. (1986)[1], Hall and Hosseini-Nasab (2006)[8], Chen and Lei (2015)[2], and Zhou et al. (2022)[20], while functional regression has been significantly refined through the work of Ramsay and Silverman (2005)[16] and Hall and Horowitz (2007)[9]. Comprehensive discussions on these topics can be found in Ramsay and Silverman (2005)[16], Hsing and Eubank (2015)[10], and Wang

et al. (2016)[18].

This study aims to develop a dynamic model based on stochastic differential equations for analyzing sparse longitudinal data by introducing a nonparametric methodology to reconstruct underlying stochastic processes without relying on strong assumptions. It seeks to enhance the accuracy of data analysis and prediction of future trajectories while comparing its performance with traditional methods such as functional principal component analysis. Through simulation studies and real-world applications, the research aims to provide a practical model applicable to various fields, including biology, economics, and social sciences, contributing to a more precise and flexible analysis of dynamic phenomena.

In recent years, stochastic differential equations have gained increasing attention as a tool for modeling dynamic stochastic phenomena. Notable contributions include the work of Dawson and Müller (2018)[4], which introduced a dynamic framework for analyzing sparse data, as well as studies exploring the estimation of SDE parameters using incomplete functional data.

Despite significant advancements in longitudinal data analysis, major challenges remain in dealing with sparse data. Some key research questions this study aims to address include: How can the underlying stochastic processes in sparse data be reconstructed using stochastic differential equations? What are the most efficient methods for estimating drift and diffusion parameters without relying on strong assumptions?

This thesis consists of three main chapters, each addressing a key aspect of dynamic modeling for sparse longitudinal data using stochastic differential equations.

Chapter 1 (Stochastic Differential Equations and Alternative Formulation):

This chapter presents the theoretical foundation of stochastic differential equations (SDEs) as a powerful tool for modeling random dynamic systems. It covers funda-

mental definitions, existence and uniqueness theorems, and discusses both classical and alternative formulations of SDEs. The chapter emphasizes the relevance of SDEs in representing time-continuous stochastic processes, particularly in the context of sparse longitudinal data.

Chapter 2 (Estimation of Stochastic Differential Equations):

This chapter focuses on estimation techniques for SDEs based on the alternative formulation introduced earlier. It explains how to simulate sample paths from sparse data and estimate conditional means and variances using nonparametric regression. The chapter also addresses convergence properties of the estimators and provides a recursive framework for constructing paths that align with the statistical properties of the underlying stochastic process.

Chapter 3 (Data Applications):

The final chapter applies the proposed methodology to sparse longitudinal data analysis, featuring a case study on child growth in Nepal using irregular height measurements. It discusses theoretical solutions for SDE based models, conducts empirical studies, compares models such as Ho-Lee and Ornstein–Uhlenbeck, and evaluates the proposed model's performance against covariance completion methods.

The study also paves the way for developing nonlinear and non-Gaussian models, improving estimation algorithms, and extending applications to fields like economics and epidemiology....

Chapter 1

Stochastic Differential Equations and an Alternative Formulation

In this chapter, we present the theoretical foundation of stochastic differential equations (SDEs), which are used to model dynamic systems influenced by randomness. We focus on the conditions for the existence and uniqueness of solutions, and introduce the alternative formulation of SDEs, which enables a more flexible analysis of sparse longitudinal data, and we mention some of the references we used [5, 13, 14, 21].

1.1 Stochastic Differential Equations

1.1.1 Motivation and General Introduction

Stochastic differential equations aim to formulate a mathematical model for a differential equation subjected to random noise. Let's look at an ordinary differential

equation given by the expression

$$X' = b(X(t)),$$

or, in differential form:

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t,$$

an equation of that kind defines the evolution of the physical system. If any random fluctuations are considered, a noise term in the form of $\sigma \mathrm{d}B_t$ is included, where B_t is Brownian Motion and σ is a constant for the time being which corresponds to the noise level. This generates a "stochastic" differential equation of the type:

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma \mathrm{d}B_t.$$

Or, in the only sense that carries any mathematical worth, the integral form:

$$X_t = x_0 + \int_0^t b(X_s)\mathrm{d}s + \sigma B_t.$$

This equation is generalized by allowing σ to depend on the state of the system at time t:

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t,$$

or, in integral form:

$$X_t = x_0 + \int_0^t b(X_s)\mathrm{d}s + \int_0^t \sigma(X_s)B_s.$$

In general, determining the solution of an SDE is not trivial. This is the case when

functions b and σ are required to meet certain conditions in order to guarantee both the existence and uniqueness of the solution of the SDE. In order to be completely general, we suppose b and σ are functions of time t. So, we investigate the following SDE:

$$X_t = X_0 + \int_0^t b(s, X_s)\mathrm{d}s + \int_0^t \sigma(s, X_s)B_s.$$

Definition 1.1.1 *A stochastic differential equation (SDE) takes the form:*

$$X_t = x_0 + \int_0^t b(s, X_s)\mathrm{d}s + \int_0^t \sigma(s, X_s)\mathrm{d}B_s,\ t \in [0, T]\ ,\ s \in [0, t], \tag{1.1}$$

or in differential form:

$$\begin{cases} \mathrm{d}X_t = b(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}B_t, \quad \forall (0 \leq t \leq T), \\ X_0 = x_0. \end{cases} \tag{1.2}$$

Let $(X_t)_{t\in[0,T]}$ be a stochastic process defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$, where T is a strictly positive number. The functions b and σ represent the drift and diffusion coefficients, respectively, and B_t denotes a Brownian motion (also referred to as a Wiener process). The initial value x_0 can either be deterministic or random, provided it is independent of the Brownian motion B_t. The filtration is defined as: $\mathcal{F}_t^{X_0} = \sigma(x_0) \vee \sigma\{B_s : 0 \leq s \leq t\}$ which represents the smallest σ-algebra generated by the initial condition X_0 and the Brownian motion B_s up to time t.

Definition 1.1.2 (Solution of an SDE)

Consider the two measurable and bounded functions $b(t, x) = \{b_i(t, x), 1 \leq i \leq n\}$ *and* $\sigma(t, x) = \{\sigma_{ij}(t, x), 1 \leq i \leq n, 1 \leq j \leq m\}$, *for all* $0 \leq t \leq T$. *Let* $n\ , m \in \mathbb{N}$, $b : \mathbb{R}^n \times [0, T] \to \mathbb{R}^n$ *and* $\sigma : \mathbb{R}^n \times [0, T] \to \mathcal{M}^{n\times m}$ *and* $x_0 \in \mathbb{R}^n$ *be an initial condition.*

A solution of SDE (1.2) is composed of:

(i) *A filtered probability space* $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ *satisfying the usual conditions.*

(ii) *An* $(\mathcal{F}_t)_{t\in[0,T]}$*-Brownian motion* $B = (B_1, \ldots, B_m)$ *with values in* $\mathbb{R}^m$.

(iii) *A process* $X = (X_t)_{t\in[0,T]}$ *continuous and* $(\mathcal{F}_t)_{t\in[0,T]}$*-adapted such that the integrals*

$$\int_0^t b(s, X_s)\mathrm{d}s \qquad \text{and} \qquad \int_0^t \sigma(s, X_s)\mathrm{d}B_s,$$

are well-defined and satisfy the equality (1.1) for $\mathbb{P}$*-a.s. for all* t.

1.2 Existence and Uniqueness Theorem.

Let $b : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \times [0,T] \to \mathcal{M}^{m\times n}$ be continuous functions that satisfy the following conditions:

a) The Lipschitz condition:

$$|b(t,x) - b(t,\widetilde{x})| + |\sigma(t,x) - \sigma(t,\widetilde{x})| \leq L\,|x - \widetilde{x}| \qquad \text{for all } 0 \leq t \leq T,\ x,\ \widetilde{x} \in \mathbb{R}^n.$$

b) The linear growth condition:

$$|b(t,x)| + |\sigma(t,x)| \leq L(1 + |x|) \qquad \text{for all } 0 \leq t \leq T,\ x \in \mathbb{R}^n.$$

for some constant L.

(c) X_0 be any random variable taking values in $\mathbb{R}^n$ such that:

$$\mathbb{E}(|X_0|^2) < \infty$$

and

$$X_0 \text{ is independent of } B^+(0).$$

Theorem 1.2.1 *Assume that (1.2) satisfies the conditions (a)(b)(c). Then, it has a unique solution* $(X_t)_{t\in[0,T]} \in \mathbb{L}^2(0,T)$.

Remark 1.2.1

1. We denote $X = (X_t)_{t\in[0,T]}$ and $\widetilde{X} = (\widetilde{X}_t)_{t\in[0,T]}$.
2. "***Unique***" means that if X, $\widetilde{X} \in \mathbb{L}^2(0,T)$, with continuous sample paths almost surely, and both solutions satisfy the stochastic differential equation (1.2), then:

$$\mathbb{P}(X_t = \widetilde{X}_t \text{ for all } 0 \leq t \leq T) = 1.$$

3. If $b(t,0)$ is bounded , hypotheses (a) claims that b and σ are uniformly Lipschitz continuous in the variable x, also note that hypothesis (b) follows from (a).

Proof. A detailed proof is presented in Evans (2013,p.91).[5] ■

1.3 An Alternative Formulation of SDEs

1.3.1 Estimation of the Drift and Diffusion Coefficients

In SDEs, the behavior of a process is mainly characterized by two functions:

- The drift, which describes the average direction of evolution.
- The diffusion, which measures the intensity of random fluctuations.

Below, we provide their mathematical definitions based on conditional expectations and variances as follows:

$$b(t,x) = \lim_{s \to t^+} \frac{1}{s-t} \mathbb{E}(X_s - X_t | X_t = x), \tag{1.3}$$

and

$$\sigma^2(t,x) = \lim_{s \to t^+} \frac{1}{s-t} \mathbb{E}((X_s - X_t)^2 | X_t = x).$$

It is worth noting that the diffusion coefficient can alternatively be defined as:

$$\sigma^2(t,x) = \lim_{s \to t^+} \frac{1}{s-t} var(X_s - X_t | X_t = x). \tag{1.4}$$

Here, $b(t, X_t)$ can be considered as the instantaneous derivative of the mean of the process conditioned on the value of $(X_t)_{t\in[0,T]}$, and $\sigma^2(t, X_t)$ can be interpreted as the instantaneous derivative of the squared fluctuations of the process conditioned on $(X_t)_{t\in[0,T]}$.

Proof. Let $s > t$ and $\mathrm{d}t = s - t$ represent an infinitesimal time increment, and let $\mathrm{d}X_t = X_s - X_t$ denote the corresponding increment of the stochastic process. Based on the SDE defined in (1.2). First we approximate the drift b of the SDE (1.2)

$$X_s - X_t = \int_t^s b(u, X_u)\mathrm{d}u + \int_t^s \sigma(u, X_u)\mathrm{d}B_u.$$

Since, the infinitesimal mean is:

$$\mathbb{E}(X_s - X_t | X_t = x) = \mathbb{E}[\int_t^s b(u, X_u)\mathrm{d}u | X_t = x] + \mathbb{E}[\int_t^s \sigma(u, X_u)\mathrm{d}B_u | X_t = x].$$

Computing the first term:

$$\mathbb{E}[\int_t^s b(u, X_u)\mathrm{d}u|X_t = x].$$

Using the fact that:

$$b(u, X_u) = b(t, X_t)+(b(u, X_u)-b(t, X_t)) \qquad \text{and} \qquad \int_t^s \mathbb{E}[b(u, X_u)-b(t, X_t)|X_t = x]\mathrm{d}u = \circ(s-t),$$

we obtain:

$$\begin{aligned}\mathbb{E}[\int_t^s b(u, X_u)\mathrm{d}u|X_t = x] &= \mathbb{E}[\int_t^s b(t, X_t) + (b(u, X_u) - b(t, X_t))\mathrm{d}u|X_t = x] \qquad (1.5)\\ &= \int_t^s \mathbb{E}[b(t, X_t)|X_t = x]\mathrm{d}u + \circ(s - t)\\ &= b(t, x)(s - t) + \circ(s - t).\end{aligned}$$

Now, to compute the second term which is $\mathbb{E}[\int_t^s \sigma(u, X_u)\mathrm{d}B_u|X_t = x]$, we rewrite the term inside the expectation as follows:

$$\mathbb{E}[\int_t^s (\sigma(u, X_u) - \sigma(t, X_t))\mathrm{d}B_u|X_t = x] + \mathbb{E}[\int_t^s \sigma(t, X_t)\mathrm{d}B_u|X_t = x],$$

since the first term represents an Itô integral, which is a martingale with mean zero, we get

$$\mathbb{E}[\int_t^s \sigma(t, X_t)\mathrm{d}B_u|X_t = x] = \sigma(t, x)\mathbb{E}(B_s - B_t) = 0, \qquad (1.6)$$

the combination between (1.5) and (1.6) leads to,

$$\begin{aligned}\mathbb{E}(X_s - X_t|X_t = x) &= \mathbb{E}\{b(t,x)(s-t) + \sigma(t,x)(B_s - B_t)\} + \circ(s-t) \\ &= b(t,x)(s-t) + \circ(s-t).\end{aligned} \tag{1.7}$$

Secondly; we approximate the diffusion process. The infinitesimal variance can be expressed as:

$$\begin{aligned}\mathbb{E}\{(X_s - X_t)^2|X_t = x\} &= \mathbb{E}[\{b(t,x)(s-t) + \sigma(t,x)(B_s - B_t)\}^2] + \circ(s-t) \\ &= \sigma^2(t,x)Var(B_s - B_t) + 2b(t,x)\sigma(t,x)(s-t)\mathbb{E}(B_s - B_t) \\ &+ \mathbb{E}(b^2(t,x)(s-t)^2) + \circ(s-t) \\ &= \sigma^2(t,x)(s-t) + b^2(t,x)(s-t)^2 + \circ(s-t).\end{aligned}$$

Thus, using the approximation in (1.7)

$$\begin{aligned}Var(X_s - X_t|X_t = x) &= \mathbb{E}\{(X_s - X_t)^2|X_t = x\} - \{b(t,x)(s-t) + \circ(s-t)\}^2 \\ &= \sigma^2(t,x)(s-t) + b^2(t,x)(s-t)^2 + \circ(s-t) - b^2(t,x)(s-t)^2 \\ &= \sigma^2(t,x)(s-t) + \circ(s-t).\end{aligned}$$

Consequently, for any positive real number ε, we can find a positive real number δ such that, for all $s > t$. where $s - t < \delta$, the following holds:

$$|b(t,x) - \frac{1}{s-t}\mathbb{E}(X_s - X_t|X_t = x)| < \varepsilon.$$

Therefore, the limit

$$\lim_{s \to t^+} \frac{1}{s-t}\mathbb{E}(X_s - X_t|X_t = x),$$

exists and equals $b(t,x)$. Similarly, we can show that the limit

$$\lim_{s\to t^+}\frac{1}{s-t}\mathbb{E}((X_s-X_t)^2|X_t=x),$$

exists and equals $\sigma^2(t,x)$. ■

1.3.2 An Alternative Formulation of SDEs

The stochastic process $(X_t)_{t\in[0,T]}$ is supposed to satisfy a general SDE as defined in (1.2). In real dataset applications, the drift and diffusion coefficients in (1.2) are typically unknown. In order to recover the underlying dynamics of X_t, instead of estimating directly the drift and diffusion terms, which is difficult for functional snippet data, we replace the representations from (1.3) and (1.4) for the drift and diffusion coefficients. This leads to the following alternative formulation of the SDE.

Lemma 1.3.1 (Alternative Formulation of SDE)

$$\begin{cases} \mathrm{d}X_t = \frac{\partial}{\partial s}\mathbb{E}(X_s|X_t)|_{s=t}\mathrm{d}t + \{\frac{\partial}{\partial s}Var(X_s|X_t)|_{s=t}\mathrm{d}t\}^{\frac{1}{2}}\mathrm{d}B_t \quad t\in[0,T], \\ X_0 = x_0. \end{cases} \tag{1.8}$$

We have taken the value of s to be greater than t when we evaluate the $\mathbb{E}(X_s \mid X_t)$ and $Var(X_s \mid X_t)$ w.r.t s. The diffusion coefficient is well-defined and nonzero. SDE (1.8) serves as the primary tool for generating sample paths of $(X_t)_{t\in[0,T]}$, given an initial condition, recursively. Given the Gaussian assumption, the distribution of $(X_t)_{t\in[0,T]}$ is constructed at every step using the conditional mean estimates $\mathbb{E}(X_s \mid X_t)$ and conditional variances $Var(X_s \mid X_t)$.

Remark 1.3.1 *we introduce some essential conditions that ensure the validity of the alternative formulation of stochastic differential equations:*

- *(A1) The mean function* $\mu(t) = \mathbb{E}(X_t)$ *is continuously differentiable on* $[0, T]$.
- *(A2) The covariance function* $\sum(s,t) = Cov(X_s, X_t)$ *is continuously differentiable in the lower triangular region* $\{(s,t) : s \geq t,\ s,\ t \in [0,T]\}$. *In other words, the partial derivatives exist and remain*

$$\sum\nolimits_s^{'}(s,t) = \frac{\partial \sum(s,t)}{\partial s}, \quad \sum\nolimits_t^{'}(s,t) = \frac{\partial \sum(s,t)}{\partial t}$$

continuous for all $s, t \in [0,T]$, *where* $s \geq t$.

- *(A3) The variance function* $\sum(t,t)$ *remains strictly positive over the half-open interval* $(0,1]$.

1.4 Existence and Uniqueness Result

Theorem 1.4.1 *If* $(X_t)_{t\in[0,T]}$ *is a Gaussian stochastic process and satisfies conditions (A1) and (A2), where the initial value* x_0 *is a random variable independent of the* σ*-algebra* $\mathcal{F}_\infty$ *generated by* $\{B_s,\ s \geq 0\}$ *and satisfying* $\mathbb{E}(x_0^2) < \infty$, *then the unique pathwise solution to the stochastic differential equation (1.8) is given by:*

$$X_t = x_0 + \int_0^t \frac{\partial}{\partial r}\mathbb{E}(X_r|X_s)|_{r=s}\mathrm{d}s + \int_0^t \{\frac{\partial}{\partial r}Var(X_r|X_s)|_{r=s}\mathrm{d}t\}^{\frac{1}{2}}\mathrm{d}B_s, \quad t \in [0,T].$$

This solution satisfies the following properties:

1. *Adaptation to Filtration:*

$$X_t \text{ is measurable w.r.t } \mathcal{F}_t^{x_0} = \sigma(x_0) \vee \sigma(B_s,\ s \in [0,t]). \tag{1.9}$$

2. *Bounded Second Moment:*

$$\sup_{[0,T]} \mathbb{E}(X_t^2) < \infty. \tag{1.10}$$

Furthermore, uniqueness holds in the sense that if two processes $(X_t)_{t\in[0,T]}$ *and* $(Y_t)_{t\in[0,T]}$ *satisfy equations (1.8), (1.9), and (1.10), then:*

$$X_t = Y_t \quad \textit{for all } t \in [0,T] \qquad \textit{a.s.}$$

Remark 1.4.1 *Given the Brownian motion* B_t *is predefined, the solution* $(X_t)_{t\in[0,T]}$ *is* $\mathcal{F}_t^{x_0}$*-adapted. where the drift coefficient* $b(t, X_t) = a(t)X_t + c(t)$ *and the diffusion coefficient is additive, i.e.,* $\sigma(t, X_t) = \sigma(t)$*. Indeed, restricting to Gaussianity,* $(X_t)_{t\in[0,T]}$ *satisfies a narrow-sense linear SDE (Kloeden and Platen 1999[11]), where the drift and diffusion coefficients are of the form:*

$$b(t, X_t) = \mu'(t) + \sum\nolimits_s^{'}(s,t)|_{s=t} \sum\nolimits^{-1}(t,t)\{X_t - \mu(t)\},$$
$$\sigma(t, X_t) = \{\sum\nolimits^{'}(t,t) - 2\sum\nolimits_s^{'}(s,t)|_{s=t}\}^{\frac{1}{2}},$$

indicating that SDE (1.8) is a narrow-sense linear. The general solution for a linear SDE can be explicitly written as:

$$X_t = \phi(t)\left\{x_0 + \int_0^t c(s)\phi^{-1}(s)\mathrm{d}s + \int_0^t \sigma(s)\phi^{-1}(s)\mathrm{d}B_s\right\},$$

where the parameters are:

$$
\begin{aligned}
a(t) &= \sum\nolimits_{s}^{'}(s,t)|_{s=t} \sum\nolimits^{-1}(t,t), \\
c(t) &= \mu'(t) - \sum\nolimits_{s}^{'}(s,t)|_{s=t} \sum\nolimits^{-1}(t,t)\mu(t), \\
\phi(t) &= e^{\int_0^t a(s)\mathrm{d}s}.
\end{aligned}
$$

A key property of the recursion in (2.1) is that it provides an exact simulation of $(X_t)_{t\in[0,T]}$ *at discrete time points* $t_1, \ldots, t_k$.

Proof. Let

$$
\begin{aligned}
b(t, X_t) &= \frac{\partial}{\partial s}\mathbb{E}(X_s|X_t)|_{s=t} \\
\sigma(t, X_t) &= \{\frac{\partial}{\partial s} Var(X_s|X_t)|_{s=t} dt\}^{\frac{1}{2}}.
\end{aligned}
$$

As stated in Theorem 1.2.1, it is sufficient to demonstrate that the linear growth and Lipschitz conditions (a) and (b) are met. Note that if $(X_t)_{t\in[0,T]}$ is a Gaussian process, then we have:

$$
\begin{aligned}
E(X_s|X_t) &= \mu(s) + \sum(s,t) \sum\nolimits^{-1}(t,t)\{X_t - \mu(t)\}, \\
Var(X_s|X_t) &= \sum(s,s) - \sum(s,t) \sum\nolimits^{-1}(t,t) \sum(t,s).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
b(t, X_t) &= (\mu'(s) + \sum\nolimits_{s}^{'}(s,t) \sum\nolimits^{-1}(t,t)\{X_t - \mu(t)\})|_{s=t} \\
&= \mu'(t) + \sum\nolimits_{s}^{'}(s,t)|_{s=t} \sum\nolimits^{-1}(t,t)\{X_t - \mu(t)\},
\end{aligned}
$$

$$\sigma(t, X_t) = [\{\sum\nolimits^{'}(s,s) - 2\sum\nolimits_{s}^{'}(s,t)\sum\nolimits^{-1}(t,t)\sum(t,s)\}|_{s=t}]^{\frac{1}{2}}$$
$$= \{\sum\nolimits^{'}(t,t) - 2\sum\nolimits_{s}^{'}(s,t)|_{s=t}\}^{\frac{1}{2}}.$$

Since $[0,T]$ is compact $\mu(t), \mu^{'}(t), \sum_{s}^{'}(s,t)$, and $\sum_{s}^{'}(t,t)$ are bounded under conditions (A1) and (A2). Additionally, if $\sum(t,t) = 0$, then $\sum(s,t) = 0$ for all $s \in [0,T]$. Consequently, $\sum(t,t) = 0$ implies that $\sum_{s}^{'}(s,t) = 0$ for all $s \in [0,T]$, and thus $\sum_{s}^{'}(s,t)|_{s=t}$ $\sum^{'}(t,t)$ remains bounded. For all $x \in \mathbb{R}$ and $t \in [0,T]$, it follows that:

$$|b(t,x)| + |\sigma(t,x)| = \left|\mu'(t) + \sum\nolimits_{s}^{'}(s,t)|_{s=t}\sum\nolimits^{-1}(t,t)\{x - \mu(t)\}\right| + \left|\{\sum\nolimits^{'}(t,t) - 2\sum\nolimits_{s}^{'}(s,t)|_{s=t}\}^{\frac{1}{2}}\right|$$
$$\leq |\mu'(t)| + \left|\sum\nolimits_{s}^{'}(s,t)|_{s=t}\sum\nolimits^{-1}(t,t)\{x - \mu(t)\}\right| + \left|\{\sum\nolimits^{'}(t,t) - 2\sum\nolimits_{s}^{'}(s,t)|_{s=t}\}^{\frac{1}{2}}\right|$$
$$\leq C_1 + C_2\,|x| + C_3$$
$$\leq C(1 + |x|).$$

For some constant C, the growth condition is therefore satisfied. Regarding the Lipschitz condition, observe that $|\sigma(t,x) - \sigma(t,y)| = 0$, we have:

$$|b(t,x) - b(t,y)| + |\sigma(t,x) - \sigma(t,y)| = |b(t,x) - b(t,y)|$$
$$= \left|\sum\nolimits_{s}^{'}(s,t)|_{s=t}\sum\nolimits^{-1}(t,t)(x - y)\right|$$
$$= \left|\sum\nolimits_{s}^{'}(s,t)|_{s=t}\sum\nolimits^{-1}(t,t)\right||x - y|$$
$$\leq C\,|x - y|$$

for all $x, y \in \mathbb{R}$, $t \in [0,T]$. ■

In conclusion, this chapter introduced the fundamental concepts of stochastic differential equations and the conditions for the existence and uniqueness of their solutions.

We also explored the alternative formulation used for handling sparse longitudinal data. However, the practical application of these models requires parameter estimation and path simulation topics that will be discussed in detail in the following chapter.

Chapter 2

Estimation of Stochastic Differential Equations

In this chapter, we move from theoretical foundations to practical implementation, focusing on parameter estimation methods for stochastic differential equations (SDEs). We explore techniques for simulating stochastic paths and estimating conditional means and variances, while addressing how to handle sparse longitudinal data using efficient recursive algorithms, and we mention some of the references we used [11, 20, 21].

2.1 Estimation

2.1.1 Simulating Sample Paths

To estimate sample paths of $(X_t)_{t\in[0,T]}$ from initial conditions using function snippets, the SDE from equation (1.8) is rewritten as:

$$\lim_{s\to t^+}(X_s - X_t)$$
$$= \lim_{s\to t^+}\left\{\frac{\mathbb{E}(X_s \mid X_t) - \mathbb{E}(X_t \mid X_t)}{s-t}(s-t) + \left\{\frac{Var(X_s \mid X_t) - Var(X_t \mid X_t)}{s-t}\right\}^{\frac{1}{2}}(B_s - B_t)\right\}$$

Starting with the initial condition $X_0 = x_0$, we can simulate the continuous-time process $(X_t)_{t\in[0,T]}$ at discrete time points. This involves defining a predetermined, evenly spaced time grid:

$0 \le t_0 < t_1 < \cdots < t_{k-1} < t_k \le 1$, where each interval has a length Δ. We denote the process's initial value at t_0 as X_0, and its simulated value at time t_k as X_k . The process is simulated recursively:

$$X_k - X_{k-1} = \frac{\mathbb{E}(X_k \mid X_{k-1}) - \mathbb{E}(X_{k-1} \mid X_{k-1})}{\Delta}\Delta + \left\{\frac{Var(X_k \mid X_{k-1}) - Var(X_{k-1} \mid X_{k-1})}{\Delta}\right\}^{\frac{1}{2}}(B_{t_k} - B_{t_{k-1}}).$$

Since $\mathbb{E}(X_{k-1} \mid X_{k-1}) = X_{k-1}$, $Var(X_{k-1} \mid X_{k-1}) = 0$, and $(B_{t_k} - B_{t_{k-1}})/\sqrt{\Delta} \sim \mathcal{N}(0,1)$, the recursion simplifies to:

$$X_k = \mathbb{E}(X_k \mid X_{k-1}) + \{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}}W_k, \qquad X_0 = x_0. \tag{2.1}$$

Where $W_k \sim \mathcal{N}(0,1)$ are independent for $k = 1, \ldots, K$.

The recursion in equation (2.1) provides an exact simulation of the Gaussian pro-

cess $(X_t)_{t\in[0,T]}$ at discrete time points, unlike traditional methods (Euler-Maruyama, Milstein) that introduce discretization errors. For practical use, especially in longitudinal studies, the time spacing Δ should be set based on the scheduled visit times, and the number of time points K is determined by Δ and the time interval of interest. To simulate $(X_t)_{t\in[0,T]}$, random samples must be drawn from $\mathcal{N}(\mathbb{E}(X_k \mid X_{k-1}),$ $Var(X_k \mid X_{k-1}))$. Since these parameters are unknown, they need to be estimated before simulation.

2.1.2 Estimation of Conditional Moments

The variable X_{k-1}, observed at time t_{k-1}, encapsulates both its value and the corresponding time index. To estimate the conditional mean $\mathbb{E}(X_k \mid X_{k-1})$ and the conditional variance $Var(X_k \mid X_{k-1})$, one can frame the problem as a regression task. In this context, X_k serves as the dependent variable, while the pair $(X_{k-1}, t_{k-1})^T$ acts as the predictor. Assume that each subject is measured at least at two distinct time points, T_{i1} and T_{i2}, yielding observations Y_{i1} and Y_{i2} , respectively. By defining $Z_i = (Y_{i1}, T_{i1})^T$ and assigning $Y_i = Y_{i2}$ for each $i = 1, ..., n$, we deal with collection $\{(Z_i, Y_i)\}_{i=1}^n$as n i.i.d samples from the joint distribution of the random variables (Z, Y). This setup leads to the formulation of the regression model as follows:

$$Y_i = m(Z_i) + \upsilon(Z_i)\varepsilon_i. \tag{2.2}$$

In this context, the functions $m(z) = \mathbb{E}(Y \mid Z = z)$ and $\upsilon^2(z) = Var(Y \mid Z = z)$ denote the conditional mean and variance of Y given $Z = z$, respectively. The error term ϵ_i is characterized by a zero conditional mean $\mathbb{E}(\varepsilon_i \mid Z_i) = 0$ and a unit conditional variance $Var(\varepsilon_i \mid Z_i) = 1$. The problem of estimating both the conditional mean and conditional variance has been extensively explored using parametric and nonparametric

regression models. In particular, to estimate the conditional variance, we follow the standard approach of modeling the squared residuals $\{Y_i - \hat{m}(Z_i)\}^2$ as the response variable and using Z_i as predictors. This regression-based approach allows the recursive simulation formula presented in equation (2.1) to be simplified accordingly.

$$X_k = m(Z_{k-1}) + \upsilon(Z_{k-1})W_k, \qquad X_0 = x_0, \tag{2.3}$$

where $Z_{k-1} = (X_{k-1}, t_{k-1})^T$ for $k = 1, ..., K$. Given estimates $\hat{m}(\cdot)$ and $\hat{\upsilon}^2(\cdot)$, the estimated sample path of $(X_t)_{t\in[0,T]}$ at $t_1, \ldots, t_k$ is computed recursively as

$$\begin{aligned} \hat{X}_1 &= \hat{m}(Z_0) + \hat{\upsilon}(Z_0)W_1, \\ \hat{X}_k &= \hat{m}(\hat{Z}_{k-1}) + \hat{\upsilon}(\hat{Z}_{k-1})W_k, \qquad k = 2, ..., K, \end{aligned} \tag{2.4}$$

where $Z_0 = (x_0, t_0)^T$ and $\hat{Z}_{k-1} = (\hat{X}_{k-1}, t_{k-1})^T$ for $k = 2, ..., K$.

In cases where the conditional mean and variance structures are unknown, non-parametric methods provide a flexible alternative to parametric approaches like multiple linear regression, though they may have a lower rate of convergence.

2.2 Some Convergence Results

In this section, we confirm that the stochastic differential equation (SDE) presented in equation (1.8) possesses a unique solution and examine the rate at which the estimated sample paths converge. The assurance of existence and uniqueness is grounded in the Gaussian nature of the process $(X_t)_{t\in[0,T]}$; that is, for any finite collection of time points $t_1, \ldots, t_k \in [0, T]$, the vector $(X_{t_1}, \ldots, X_{t_k})^T$ exhibits a joint Gaussian distribution. A pivotal aspect of our analysis involves articulating the conditional expectation $\mathbb{E}(X_s \mid X_t)$ and the conditional variance $Var(X_s \mid X_t)$ using the mean and covariance

functions associated with $(X_t)_{t\in[0,T]}$. This approach facilitates the demonstration that the drift and diffusion coefficients specified in equation(1.8) adhere to the Lipschitz and linear growth conditions from (a) and (b):

$$\mathbb{E}(X_s|X_t) = \mu(s) + \sum(s,t)\sum^{-1}(t,t)\{X_t - \mu(t)\},$$

$$Var(X_s|X_t) = \sum(s,s) - \sum(s,t)\sum^{-1}(t,t)\sum(t,s).$$

Even when dealing with a non-Gaussian process $(X_t)_{t\in[0,T]}$, as long as a unique solution is present, it is possible to determine the convergence rate of the estimated sample path, provided that both the conditional mean $m(\cdot)$ and conditional variance functions $v^2(\cdot)$ adhere to the Lipschitz continuity condition.

Remark 2.2.1 *To ensure that the drift and diffusion coefficients in (1.8) meet the Lipschitz and linear growth conditions (a) and (b), the conditions (A1) and (A2) are required. These conditions impose regularity constraints on the process* $(X_t)_{t\in[0,T]}$. *In particular, (A2) ensures that the covariance function* $\sum(s,t)$ *is continuously differentiable in the upper triangular region*$\{(s,t) : s \leq t,\ s,\ t \in [0,T]\}$, *though it may not be differentiable along the diagonal* $s = t$. *This non-differentiability at* $s = t$ *is a well-known property, as seen in Brownian motion, where the covariance function* $\sum(s,t) = \min(s,t)$ *is continuous but not differentiable along the diagonal.*

Lemma 2.2.1 *States that if the stochastic process* $(X_t)_{t\in[0,T]}$ *follows a Gaussian distribution, the recursion in (2.1) provides an exact simulation of the process at time points* $t_1, \ldots, t_k$, *ensuring that the generated values match the true distribution of the continuous-time process at these points. Additionally, this Lemma guarantees that discretization errors are not a concern when analyzing the convergence rate of the estimated*

sample path if $(X_t)_{t\in[0,T]}$ *is Gaussian. The asymptotic properties of the estimated path are examined based on (2.4), focusing on the rate of convergence for* $\hat{X}_K$*, which also applies to* $\hat{X}_k$ *for any* k*. The proof relies on a recursive formula for the difference* $\left|\hat{X}_k - X_k\right|$ *and utilizes the Lipschitz continuity of the conditional mean function* $m(\cdot)$ *and the conditional variance function* $\upsilon^2(\cdot)$ *. To establish this, we require the conditions (A3) regarding the variance function* $\sum(t,t)$ *and the design of functional snippets. This assumption is reasonable in practical applications, particularly when modeling the stochastic dynamics of* $(X_t)_{t\in[0,T]}$*. Under Gaussianity, the recursive expressions for the conditional mean and variance in (2.3) simplify to:*

$$m(Z_{k-1}) = \mu(t_k) + \sum(t_k, t_{k-1}) \sum^{-1}(t_{k-1}, t_{k-1})\{X_{t_{k-1}} - \mu(t_{k-1}))\}, \tag{2.5}$$

$$\upsilon^2(Z_{k-1}) = \sum(t_k, t_k) - \sum(t_k, t_{k-1}) \sum^{-1}(t_{k-1}, t_{k-1}) \sum(t_{k-1}, t_k), \tag{2.6}$$

where $Z_{k-1} = (X_{k-1} - t_{k-1})^T$ *and* $t_k = t_{k-1} + \Delta$ *represents the discrete time points used to simulate the sample path of the underlying process* $(X_t)_{t\in[0,T]}$*.*

Proof. Consider a recursive relationship (2.1) defined for $k = 1, \ldots, K$ as follows:

$$X_k = \mathbb{E}(X_k \mid X_{k-1}) + \{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k.$$

Here, $\mathbb{E}(X_k \mid X_{k-1})$ represents the conditional expectation of X_k given X_{k-1}, while $Var(X_k \mid X_{k-1})$ represents the corresponding conditional variance. W_k denotes a standard Gaussian random variable. Assuming that X_k follows a Gaussian distribution, it is sufficient to demonstrate that both the mean and variance of the right-hand side of

this recursion match $\mathbb{E}(X_k)$ and $Var(X_k)$, respectively. To prove the mean:

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}(X_k \mid X_{k-1}) + \{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k] \\
&\qquad = \mathbb{E}[\mathbb{E}(X_k \mid X_{k-1})] + \mathbb{E}[\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k] \\
&\qquad = \mathbb{E}(X_k) + \mathbb{E}[\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}}].\mathbb{E}(W_k) \\
&\qquad = \mathbb{E}(X_k).
\end{aligned}$$

This holds because $W_k \sim \mathcal{N}(0,1)$ and it is independent . To prove the variance: Given that W_k is independent of both $\mathbb{E}(X_k \mid X_{k-1})$ and $\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}}$, we have:

$$\begin{aligned}
&Var[\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k] \\
&\qquad = \mathbb{E}[\{Var(X_k \mid X_{k-1})\} W_k^2] - \mathbb{E}^2[\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k] \\
&\qquad = \mathbb{E}[\{Var(X_k \mid X_{k-1})]\mathbb{E}[W_k^2] - \mathbb{E}^2[\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}}]\mathbb{E}^2[W_k] \\
&\qquad = \mathbb{E}\{Var(X_k \mid X_{k-1})\}.
\end{aligned}$$

Also, the covariance is:

$$\begin{aligned}
&cov[\mathbb{E}(X_k \mid X_{k-1}), \{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k] \\
&\qquad = \Big\{\mathbb{E}[\mathbb{E}(X_k \mid X_{k-1}).\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k] \\
&\qquad -\mathbb{E}\{\mathbb{E}(X_k \mid X_{k-1})\}.\mathbb{E}[\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k]\Big\} \\
&\qquad = \Big\{\mathbb{E}[\mathbb{E}(X_k \mid X_{k-1}).\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}}].\mathbb{E}(W_k) \\
&\qquad\quad -\mathbb{E}(X_k).\mathbb{E}[\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}}].\mathbb{E}(W_k)\Big\} \\
&\qquad = 0
\end{aligned}$$

Thus, the total variance is:

$$\begin{aligned} &Var[\mathbb{E}(X_k \mid X_{k-1}) + \{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k] \\ &\quad = \Big\{ Var\{\mathbb{E}(X_k \mid X_{k-1})\} + Var[\{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k] \\ &\qquad +2cov[\mathbb{E}(X_k \mid X_{k-1}), \{Var(X_k \mid X_{k-1})\}^{\frac{1}{2}} W_k] \Big\} \\ &\quad = Var\{\mathbb{E}(X_k \mid X_{k-1})\} + \mathbb{E}\{Var(X_k \mid X_{k-1})\} \\ &\quad = Var(X_k) \end{aligned}$$

■

Lemma 2.2.2 *States that if the stochastic process $(X_t)_{t\in[0,T]}$ is Gaussian and satisfies conditions (A1), (A2), and (A3), then for $k = 2, \ldots, K$, the conditional mean and conditional variance in recursion (2.3) satisfy:*

$$\begin{aligned} \left| m(\hat{Z}_{k-1}) - m(Z_{k-1}) \right| &\leq L \left| \hat{X}_{k-1} - X_{k-1} \right|, \\ \left| \upsilon(\hat{Z}_{k-1}) - \upsilon(Z_{k-1}) \right| &= 0, \end{aligned}$$

where $L = \max_{t\in\{t_1,\ldots,t_{K-1}\}} \left| \sum(t+\Delta, t) \sum^{-1}(t,t) \right|$ *and* $Z_{k-1} = (X_{k-1}, t_{k-1})^T$, $\hat{Z}_{k-1} = (\hat{X}_{k-1}, t_{k-1})^T$.

This lemma ensures that the sequence $\left| \hat{X}_k - X_k \right|$ *does not increase rapidly, allowing one to bound* $\left| \hat{X}_k - X_k \right|$ *recursively. To determine the convergence rate of the estimated sample path, one must analyze the asymptotic behavior of the estimates of the conditional mean function $\hat{m}(.)$ and the conditional variance function $\upsilon^2(\cdot)$. Suppose that for any fixed $z \in \mathbb{R} \times [0,T]$, the following holds:*

$$\begin{aligned} [\mathbb{E}\{ \left| \hat{m}(z) - m(z) \right|^2 \}]^{\frac{1}{2}} &= O(\alpha_n) \\ [\mathbb{E}\{ \left| \hat{\upsilon}^2(z) - \upsilon^2(z) \right|^2 \}]^{\frac{1}{2}} &= O(\beta_n). \end{aligned} \tag{2.7}$$

If the residual-based estimator is used to estimate the conditional variance function $v^2(\cdot)$*, it is well known that estimating the conditional mean function* $\hat{m}(.)$ *does not affect the estimation of* $v^2(\cdot)$ *(Fan and Yao 1998[6]). Therefore, if the same regression method is used for both* $m(.)$ *and* $v^2(\cdot)$*, then,* $\alpha_n = \beta_n$*. For multiple linear regression,* $\alpha_n = \beta_n = n^{-\frac{1}{2}}$*, whereas for local linear regression,* $\alpha_n = \beta_n = n^{-\frac{1}{3}}$*.*

Proof. Under the assumption of Gaussianity, the conditional mean and conditional variance in recursion (2.3) simplify to:(2.5,2.6), where $Z_{k-1} = (X_{k-1} - t_{k-1})^T$ and $t_k = t_{k-1} + \Delta$ represents the discrete time points used to simulate the sample path of the underlying process $(X_t)_{t\in[0,T]}$.

Since $[0, T]$ is compact, it follows from Conditions (A1) and (A2) that $\mu(t)$ and $\sum(s,t)$ are bounded. Additionally, for $k = 2, \ldots, K$, the term $\sum(t_k, t_{k-1}) \sum^{-1}(t_{k-1}, t_{k-1})$ is bounded under Condition (A3), as $t_{k-1} > 0$ for $k = 2, \ldots, K$ and $t_k = t_{k-1} + \Delta$.

Now, defining $\hat{Z}_{k-1} = (\hat{X}_{k-1}, t_{k-1})^T$, for $k = 2, \ldots, K$, we obtain:

$$\begin{aligned}
\left|m(\hat{Z}_{k-1}) - m(Z_{k-1})\right| &= \Big|(\mu(t_k) + \sum(t_k, t_{k-1}) \sum^{-1}(t_{k-1}, t_{k-1})\{\hat{X}_{t_{k-1}} - \mu(t_{k-1}))\}) \\
&\quad -(\mu(t_k) + \sum(t_k, t_{k-1}) \sum^{-1}(t_{k-1}, t_{k-1})\{X_{t_{k-1}} - \mu(t_{k-1}))\})\Big| \\
&= \left|\sum(t_k, t_{k-1}) \sum^{-1}(t_{k-1}, t_{k-1})(\hat{X}_{k-1} - X_{k-1})\right| \\
&\leq L \left|\hat{X}_{k-1} - X_{k-1}\right|.
\end{aligned}$$

where $L = \max_{t\in\{t_1,\ldots,t_{K-1}\}} \left|\sum(t+\Delta, t) \sum^{-1}(t,t)\right|$. Since, under the Gaussian assumption, the conditional variance $v^2(Z_{k-1})$ depends only on t_{k-1}, it follows directly that:

$$\begin{aligned}
\left|v(\hat{Z}_{k-1}) - v(Z_{k-1})\right| &= \Big|(\sum(t_k, t_k) - \sum(t_k, t_{k-1}) \sum^{-1}(t_{k-1}, t_{k-1}) \sum(t_{k-1}, t_k)\})^{\frac{1}{2}} \\
&\quad -(\sum(t_k, t_k) - \sum(t_k, t_{k-1}) \sum^{-1}(t_{k-1}, t_{k-1}) \sum(t_{k-1}, t_k)\})^{\frac{1}{2}}\Big| \\
&= 0
\end{aligned}$$

■

Theorem 2.2.1 *If the stochastic process $(X_t)_{t\in[0,T]}$ is Gaussian and satisfies conditions (A1), (A2), and (A3), then for the estimated sample path of the stochastic differential equation (SDE) (1.8), as defined in (2.4), the following holds:*

$$\{\mathbb{E}(\left|\hat{X}_k - X_k\right|^2)\}^{\frac{1}{2}} = O(\alpha_n + \beta_n),$$

where α_n and β_n represent the convergence rates of the conditional mean function estimate $\hat{m}(.)$ and the conditional variance function estimate $v^2(.)$, as given in (2.7). This theorem establishes that $\hat{X}_K$ strongly converges to X_K, ensuring that both the mean and variance converge, specifically:

$$\left|\mathbb{E}(\hat{X}_K) - \mathbb{E}(X_K)\right| = O(\alpha_n + \beta_n)$$
$$\left|var(\hat{X}_K) - var(X_K)\right| = O(\alpha_n^2 + \beta_n^2).$$

This convergence holds uniformly over k, confirming the pathwise convergence of the estimated sample path to the true process. Denoting the distributions of X_K and its corresponding estimator $\hat{X}_K$ as $\mathcal{L}(X_K)$ and $\mathcal{L}(\hat{X}_K)$, respectively, we aim to measure the discrepancy between these distributions as an indicator of the estimator's performance. The strong convergence results from Theorem 2 allow us to determine the convergence rate of the 2-Wasserstein distance $\mathrm{d}W = \{\mathcal{L}(\hat{X}_K), \mathcal{L}(X_K)\}$, see[17]. The 2-Wasserstein distance between two probability measures v_1 and v_2 on $\mathbb{R}$ is given by

$$\mathrm{d}_W^2(v_1, v_2) = \int_0^1 \left\{F_1^{-1}(p) - F_2^{-1}(p)\right\}^2 \mathrm{d}p,$$

where F_1^{-1} and F_2^{-1} represent the quantile functions of v_1 and v_2, respectively. In the case where are v_1 and v_2 one-dimensional Gaussian distributions with means and variances

(m_1, σ_1^2) and (m_2, σ_2^2) , the squared 2-Wasserstein distance simplifies to

$$\mathrm{d}_W^2(v_1, v_2) = (m_1 - m_2)^2 + (\sigma_1 - \sigma_2)^2.$$

Then, we derive the rate of convergence for the Wasserstein distance.

Proof. Observe that the conditional variance

$$Var(X_s|X_t) = \sum(s,s) - \sum(s,t)\sum^{-1}(t,t)\sum(t,s), \quad 0 \le t < s \le 1,$$

is always nonnegative and equals zero , meaning $Var(X_s|X_t) \geq 0$ if and only if :

- X_s is deterministic
- or $X_s = cX_t$ for some constant $c \neq 0$.

The first case is ruled out by Condition (A3), while the second case is prevented due to the presence of $\mathrm{d}B_t$ in the stochastic differential equation (SDE) governing $(X_t)_{t\in[0,T]}$. As a result, $Var(X_s|X_t)$ remains strictly positive for all $0 \le t < s \le 1$. Consequently, the conditional variance function

$$v^2(z) = \sum(t+\delta, t+\delta) - \sum(t+\delta, t)\sum^{-1}(t,t)\sum(t, t+\delta),$$

is also bounded away from zero for any fixed $z = (x,t)^T$ and $t \in [0, 1-\delta]$. From equation (2.7), it follows that

$$\mathbb{E}\{\left|\hat{v}^2(z) - v^2(z)\right|^2\} = O(\beta_n^2).$$

For any fixed $z = (x,t)^T$ with $t \in [0, 1-\delta]$, using $a - b = \frac{a^2-b^2}{a+b}$, the following holds:

$$\begin{aligned}
\mathbb{E}\{\,|\hat{\upsilon}(z) - \upsilon(z)|^2\,\} &= \mathbb{E}\left\{\frac{|\hat{\upsilon}^2(z) - \upsilon^2(z)|^2}{|\hat{\upsilon}(z) + \upsilon(z)|^2}\right\} \\
&\leq \mathbb{E}\left\{\frac{|\hat{\upsilon}^2(z) - \upsilon^2(z)|^2}{\upsilon^2(z)}\right\} \\
&= \frac{1}{\upsilon^2(z)}\mathbb{E}\left\{\left|\hat{\upsilon}^2(z) - \upsilon^2(z)\right|^2\right\} \\
&= O(\beta_n^2). \qquad (2.8)
\end{aligned}$$

Since $\hat{\upsilon}(z) \geq 0$ and $\upsilon^2(z)$ is strictly bounded away from zero, we recall the recursive procedure

$$\begin{aligned}
\hat{X}_1 &= \hat{m}(Z_0) + \hat{\upsilon}(Z_0)W_1, \\
\hat{X}_k &= \hat{m}(\hat{Z}_{k-1}) + \hat{\upsilon}(\hat{Z}_{k-1})W_k, \qquad k = 2, ..., K,
\end{aligned}$$

where $Z_0 = (x_0, t_0)^T$ and $\hat{Z}_{k-1} = (\hat{X}_{k-1}, t_{k-1})^T$ for $k = 2, ..., K$. For $k = 1$, it follows that

$$\begin{aligned}
\mathbb{E}(\left|\hat{X}_1 - X_1\right|^2) &= \mathbb{E}[\,|\hat{m}(Z_0) - m(Z_0) + \{\hat{\upsilon}(Z_0) - \upsilon(Z_0)\}W_1|^2\,] \\
&= \mathbb{E}\{\,|\hat{m}(Z_0) - m(Z_0)|^2\} + \mathbb{E}[\,|\{\hat{\upsilon}(Z_0) - \upsilon(Z_0)\}W_1|^2] \\
&+ 2\mathbb{E}(\{\hat{m}(Z_0) - m(Z_0)\}[\{\hat{\upsilon}(Z_0) - \upsilon(Z_0)\}W_1]) \\
&= \mathbb{E}\{\,|\hat{m}(Z_0) - m(Z_0)|^2\} + \mathbb{E}\{\,|\hat{\upsilon}(Z_0) - \upsilon(Z_0)|^2\}\mathbb{E}(W_1^2) \\
&= \mathbb{E}\{\,|\hat{m}(Z_0) - m(Z_0)|^2\} + \mathbb{E}\{\,|\hat{\upsilon}(Z_0) - \upsilon(Z_0)|^2\} \\
&= O(\alpha_n^2 + \beta_n^2). \qquad (2.9)
\end{aligned}$$

Since $W_1 \sim \mathcal{N}(0,1)$ is independent of $\hat{m}(Z_0)$ and $\hat{\upsilon}(Z_0)$, we can extend this result for

any $k = 2, \ldots, K$, as follows:

$$\begin{aligned}\mathbb{E}(\left|\hat{X}_k - X_k\right|^2) &= \mathbb{E}[\left|\hat{m}(\hat{Z}_{k-1}) - m(Z_{k-1}) + \{\hat{\upsilon}(\hat{Z}_{k-1}) - \upsilon(Z_{k-1})\}W_k\right|^2] \\ &= \mathbb{E}\{\left|\hat{m}(\hat{Z}_{k-1}) - m(Z_{k-1})\right|^2\} + \mathbb{E}\{\left|\hat{\upsilon}(\hat{Z}_{k-1}) - \upsilon(Z_{k-1})\}\right|^2\}, \quad (2.10)\end{aligned}$$

where $Z_{k-1} = (X_{k-1} - t_{k-1})^T$, since $W_1 \sim \mathcal{N}(0,1)$ is independent of $\hat{m}(\hat{Z}_{k-1})$ and $\hat{\upsilon}(\hat{Z}_{k-1})$. Additionally, we observe that:

$$\hat{m}(\hat{Z}_{k-1}) - m(Z_{k-1}) = \{\hat{m}(\hat{Z}_{k-1}) - m(\hat{Z}_{k-1})\} + \{m(\hat{Z}_{k-1}) - m(Z_{k-1})\}.$$

The first term satisfies

$$\mathbb{E}[\,|\hat{m}(\hat{z}_{k-1}) - m(\hat{z}_{k-1})|^2\,] = O(\alpha_n^2),$$

by equation (2.7). The second term follows from Lemma 2.2.2:

$$\left|m(\hat{Z}_{k-1}) - m(Z_{k-1})\right| \leq L\left|\hat{X}_{k-1} - X_{k-1}\right|,$$

which leads to

$$\mathbb{E}[\left|m(\hat{Z}_{k-1}) - m(Z_{k-1})\right|^2] \leq L^2\mathbb{E}[\left|\hat{X}_{k-1} - X_{k-1}\right|^2].$$

Thus, we obtain

$$\begin{aligned}\mathbb{E}[\left|m(\hat{Z}_{k-1}) - m(Z_{k-1})\right|^2] &= \mathbb{E}[\left|(\hat{m}(\hat{Z}_{k-1}) - m(\hat{Z}_{k-1})) + (m(\hat{Z}_{k-1}) - m(Z_{k-1}))\right|^2] \\ &= O(\alpha_n^2 + L^2\mathbb{E}[\left|\hat{X}_{k-1} - X_{k-1}\right|^2]).\end{aligned}$$

For the second term in equation (2.10), we have

$$\hat{\upsilon}(\hat{Z}_{k-1}) + \upsilon(Z_{k-1}) = (\hat{\upsilon}(\hat{Z}_{k-1}) - \upsilon(\hat{Z}_{k-1})) + (\upsilon(\hat{Z}_{k-1}) - \upsilon(Z_{k-1})).$$

From Lemma 22.2.2, it follows that

$$\upsilon(\hat{Z}_{k-1}) - \upsilon(Z_{k-1}) = 0,$$

so

$$\hat{\upsilon}(\hat{Z}_{k-1}) - \upsilon(Z_{k-1}) = \hat{\upsilon}(\hat{Z}_{k-1}) - \upsilon(\hat{Z}_{k-1}).$$

By equation (2.8),

$$\mathbb{E}[\left|\hat{\upsilon}(\hat{Z}_{k-1}) - \upsilon(Z_{k-1})\right|^2] = \mathbb{E}[\left|\hat{\upsilon}(\hat{Z}_{k-1}) - \upsilon(\hat{Z}_{k-1})\right|^2] = O(\beta_n^2).$$

Thus, we derive

$$\mathbb{E}[\left|\hat{X}_k - X_k\right|^2] = O(L^2\mathbb{E}[\left|\hat{X}_{k-1} - X_{k-1}\right|^2] + \alpha_n^2 + \beta_n^2).$$

By applying this recursive formula along with equation (2.9), we get

$$\mathbb{E}[\left|\hat{X}_k - X_k\right|^2] = O(\alpha_n^2 + \beta_n^2),$$

for any $k = 2, \ldots, K$. Consequently,

$$(\mathbb{E}[\left|\hat{X}_k - X_k\right|^2])^{\frac{1}{2}} = O(\alpha_n + \beta_n).$$

■

In conclusion, this chapter presented various methods for estimating stochastic differential equations, including path simulation and the estimation of key parameters using nonparametric regression models. We also discussed convergence results and the accuracy of the estimated sample paths compared to the true underlying process.

However, several important questions remain: How do these methods perform when applied to real-world data? Can the proposed model truly capture the underlying trends in irregular growth data? And how does it compare to traditional statistical models? These questions will be addressed in the following chapter through real data applications, particularly in the context of child growth data from Nepal.

Chapter 3

Data Applications

This chapter focuses on the simulation process and analysis of results to evaluate the effectiveness of the proposed stochastic differential equation (SDE) models. The simulations are designed to reflect real-world data dynamics, ensuring the accuracy of estimated sample paths. The results are then analyzed to assess the performance of the estimation methods and their applicability in modeling sparse and complex longitudinal data, and it's based on the following references [3, 6, 7, 15, 17, 19, 20, 21].

3.1 Motivation and Example

3.1.1 Nepal Growth Study Data

One of the fundamental challenges in analyzing functional and longitudinal data is the growth and variability of data across different populations and study conditions. In particular, studies focusing on pediatric growth rely on sparse and irregular data collection due to practical constraints. Traditional statistical methods often struggle

to provide accurate and individualized predictions due to the missing data and the difficulty in modeling the underlying dynamics.

A prime example is the Nepal Growth Study, which tracked height measurements for 2258 children from birth to 76 months, with irregular intervals between measurements. The dataset contained between 1 and 5 observations per child, leading to fragmented growth trajectories. This variability presents a key challenge: how to infer meaningful growth patterns from incomplete and sparse data?

Problem Statement

The problem can be framed as follows:

1. Sparse Data Availability: Many children have only one or two measurements, making it hard to estimate continuous growth trends.

2. Irregular Observation Intervals: Measurements are not taken at standardized intervals, which complicates direct curve fitting approaches.

3. Sex-Based Growth Differences: Male and female growth trajectories differ significantly, requiring gender-specific modeling.

4. Need for Predictive Modeling: Traditional functional data analysis relies heavily on covariance estimation, which fails when data is sparse.

Research Approach

To address these challenges, we utilize a Stochastic Differential Equation (SDE)-based approach to model dynamic distributions. This method:

- Allows growth predictions based on limited past data.

- Uses dynamic probability models rather than direct covariance estimation.

- Provides individualized forecasts for subjects with only a single or very few observations.

The effectiveness of this method will be tested through simulations, demonstrating its capability to recover growth trends from sparse datasets.

3.1.2 Some Examples of *Alternative Formulation of* SDEs

The Ho-Lee and Ornstein–Uhlenbeck models are key examples of stochastic differential equations (SDEs) used to model dynamic systems. The Ho-Lee model is widely applied in finance to simulate interest rate movements, while the Ornstein–Uhlenbeck model is used in physics and economics to describe mean-reverting stochastic processes. Both models help capture randomness and predict system behavior under uncertainty.

Ho-Lee Model

The Ho-Lee model describes the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = g(t)\mathrm{d}t + \sigma \mathrm{d}B_t,$$

where $\sigma > 0$ and $g(t)$ is a deterministic function of time. Integrating both sides from 0 to t yields:

$$X_t = X_0 + \int_0^t g(s)\mathrm{d}s + \sigma B_t.$$

Given $X_0 = x_0$, it follows that:

$$\mathbb{E}(X_t) = x_0 + \int_0^t g(s)\mathrm{d}s,$$

and

$$Cov(X_s, X_t) = \sigma^2 \min(s, t).$$

As long as $g(t)$ is continuous, the mean function remains continuously differentiable on $[0, T]$. In the lower triangular region $\{(s, t) : s \geq t,\, s, t \in T\}$, the covariance function simplifies to $\sigma^2 t$, which is also continuously differentiable. The variance function $\sigma^2 t$ remains strictly positive over the half-open interval $(0, 1]$. Thus, we conclude that $(X_t)_{t\in[0,T]}$ satisfies the conditions for smoothness of the mean function, smoothness of the covariance function, and positivity of the variance function.

Indeed,

$$X_t = X_0 + \int_0^t g(u)\mathrm{d}u + \sigma B_t,$$
$$X_s = X_0 + \int_0^s g(u)\mathrm{d}u + \sigma B_s,$$

since $s > t$, then, $B_s = B_t + (B_s - B_t)$

$$X_s = X_0 + \int_0^t g(u)\mathrm{d}u + \sigma B_t + \int_t^s g(u)\mathrm{d}u + \sigma(B_s - B_t)$$
$$X_s = X_t + \int_t^s g(u)\mathrm{d}u + \sigma(B_s - B_t).$$

Thus,

$$\begin{aligned}\mathbb{E}(X_s|X_t) &= X_t + \int_t^s g(u)\mathrm{d}u + \sigma\mathbb{E}(B_s - B_t|X_t) \\ &= X_t + \int_t^s g(u)\mathrm{d}u.\end{aligned}$$

Then,

$$\begin{aligned}\frac{\partial}{\partial s}\mathbb{E}(X_s|X_t)|_{s=t} &= g(t) \\ &= b(t,x).\end{aligned}$$

And

$$\begin{aligned}var(X_s|X_t) &= var(\sigma(B_s - B_t)) \\ &= \sigma^2(s-t)\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\partial}{\partial s}var(X_s|X_t)|_{s=t} &= \sigma^2 \\ \{\frac{\partial}{\partial s}var(X_s|X_t)|_{s=t}\}^{\frac{1}{2}} &= \sigma.\end{aligned}$$

The Ornstein–Uhlenbeck Process

Let $\theta > 0$, $\sigma > 0$. The one-dimensional Ornstein–Uhlenbeck process satisfies the stochastic differential equation:

$$\mathrm{d}X_t = -\theta X_t\mathrm{d}t + \sigma\mathrm{d}B_t.$$

This equation is solved using Itô's formula applied to $e^{\theta t}X_t$, and we obtain:

$$X_t = X_0 e^{-\theta t} + \sigma \int_0^t e^{-\theta(t-s)} \mathrm{d}B_s.$$

Note that the stochastic integral is a Wiener integral, as the integrand is deterministic, meaning it belongs to the Gaussian space generated by B.

First, consider the case where $X_0 = x_0 \in \mathbb{R}$. $(X_t)_{t\in[0,T]}$ is a (non-centered) Gaussian process, whose mean function is $\mathbb{E}[X_t] = x_0 e^{-\theta t}$, and whose covariance function is:

$$cov(X_s, X_t) = \frac{\sigma^2}{2\theta}(e^{-\theta|t-s|} - e^{-\theta(t+s)}).$$

The mean function remains continuously differentiable on $[0, T]$. In the lower triangular region $\{(s,t) : s \geq t,\ s,\ t \in T\}$, the covariance function simplifies to $\frac{\sigma^2}{2\theta}e^{-\theta s}(e^{\theta t} - e^{-\theta t})$, which is also continuously differentiable. The variance function $\frac{\sigma^2}{2\theta}(1 - e^{-2\theta t})$ remains strictly positive over the half-open interval $(0, 1]$. Thus, we conclude that X_t satisfies the conditions for smoothness of the mean function, smoothness of the covariance function, and positivity of the variance function.

3.2 Simulation Studies

3.2.1 Analysis of Children's Growth Data in Nepal

Monitoring children's growth is crucial in public health as it helps in early detection of developmental issues. Due to limited resources in many rural areas, studies often rely on incomplete data. This study aims to analyze children's growth data in Nepal using a dynamic model capable of predicting growth patterns based on incomplete datasets. We

demonstrate the potential of the proposed dynamic modeling approach to characterize underlying growth patterns and reveal specific growth trends with snippet data from a Nepal growth study (West Jr et al.1997[19]).

Study Methodology

Height data for $2,258$ children from rural Nepal were collected at five different time points, ranging from birth to 76 months, with measurements taken approximately four months apart. To analyze these data, a subset of $1,000$ records was selected, including 107 males and 93 females. These data are extracted from a study on growth in Nepal (West Jr et al.1997[19] , Chen and Müller 2012[3]). Due to missing data, the actual number of measurements per child ranged between 1 and 5. The children with at least two measurements in a row were included in the model, while the rest were used for model validation. The model was applied to females and males separately since female and male growth trends differ significantly, with females reaching puberty far sooner than males. Up to this point the number of measurements per subject N_i was taken to be 2 for simplicity. For denser scenarios where $N_i > 2$, one could divide N_i the measurements into $N_i - 1$ pairs of contiguous measurements for each child and combine these pairs into a new sample for conditional mean and conditional variance estimation. This is a useful approach to augment the sample size, especially if the sample size n is relatively small, which is often the case in practice.

Sample Distribution The table below shows the distribution of children based on gender and the number of recorded measurements per child. The children included in the model had at least two consecutive measurements, while the rest were retained for validation purposes.

Number of Measurements per Child	Number of Males	Number of Females	Total Children
1 Measurement	30	25	55
2 Measurement	40	35	75
3 Measurement	20	15	35
4 Measurement	10	10	20
5 Measurement	7	8	15
Total	107	93	200

Table 3.1: Children's Height Distribution in the Nepal Growth Study

Handling Missing Data Since not all children had complete measurements, a dynamic model was applied to handle missing data by combining adjacent measurement pairs to estimate future growth trends. This strategy helps to make full use of all available information by transforming each set of measurements into pairs and estimating the conditional mean and variance using local linear regression techniques.

Data Analysis and Results

The model was applied separately to males and females due to significant differences in their growth patterns. Two children were selected as case studies to analyze the impact of missing data:

- **Female Child:** Had only one measurement at 4 months (52.9 cm).
- **Male Child:** Had two measurements at 12 months (63 cm) and 20 months (65.1cm).

To predict the future growth trajectories of two selected children, we implemented a recursive procedure , as detailed in Equation (2.4), 100 times. This analysis utilized

growth snippets that included at least two measurements in a row. Local linear regression was adopted to estimate the conditional mean and conditional variance. The starting time was set at $t_0 = 4$ months for the selected female and $t_0 = 12$ months for the selected male, with time intervals of $\Delta = 4$ months, corresponding to the intended measurement intervals of the Nepal growth study.

Comparison of Actual Height with Estimated Growth Curves The actual height measurements of the selected children are compared with the estimated growth curves to assess the accuracy of the predictive model. The model was run 100 times to generate future growth curves, calculating the 5%, 50%, and 95% percentiles to compare actual and predicted values. Despite the limited and fragmented nature of the available data, the proposed method effectively captures meaningful growth patterns from the observed snippets, enabling accurate predictions of future growth trends for the selected children. The following table presents the observed heights alongside the predicted values at specific age points:

Child	Age (Months)	Actual Height (cm)	Predicted Mean Mean (cm)	5th Percentile (cm)	95th Percentile (cm)
Female	4	52.9	53.0	50.5	55.2
Male	12	63.0	64.2	61.0	67.5
Male	20	65.1	68.0	66.0	70.5

Table 3.2: Growth Measurements per Child :Summary Statistics

Estimated Growth Curves for Children The figure below compares the observed and estimated growth trends for males and females, including the percentile curves and the observed growth points. For the selected male, one fresh height measurement

became available at 20 months old, which fell below the 5% percentile curve, indicating a potential developmental concern.

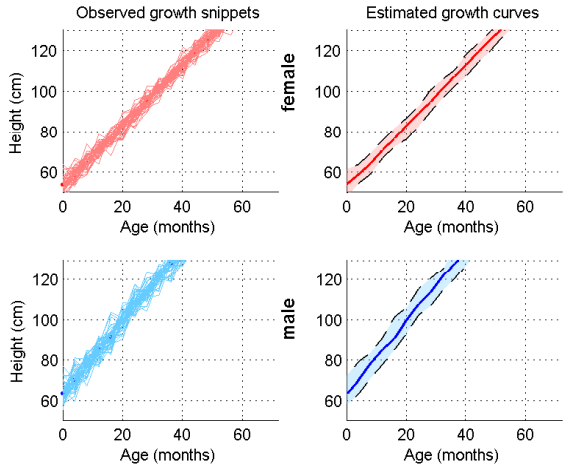

Figure 3.1: Analyzing Growth Snippets and Estimated Growth Curves by Gender in the Nepal Growth Study

Growth Assessment and Recommendations

- The analysis of height at 20 months showed that the male child's height (65.1 cm) was below the 5th percentile, indicating potential developmental delay that may require further follow-up.
- The female child's growth appeared to follow the predicted trajectory without concern.
- As children grow and new measurements are collected, we can monitor their development by comparing these fresh data points to the predicted growth trajectories.

Conclusions and Recommendations

1. Model Effectiveness: The proposed model successfully predicted growth trends even with incomplete data.

2. Importance of Monitoring: This approach can be used to track children with lower-than-expected growth rates, allowing for early intervention.

3. Enhancing Data Collection: Increasing the frequency of measurements is recommended to improve prediction accuracy and reduce estimation errors.

3.2.2 A Simulation Study of the Ho-Lee and Ornstein-Uhlenbeck Processes

Comparison Between LW and Stochastic DM Methods

In this study, we demonstrate the effectiveness of our proposed method in uncovering underlying dynamics from functional snippets across various simulation settings. Unlike traditional approaches that rely solely on covariance completion, our method offers a significant advantage: it bypasses the need for covariance estimation and does not depend on functional principal component analysis (FPCA).

For comparative purposes, we employ the covariance completion technique introduced by Lin and Wang (2022)[15], referred to as LW, utilizing the mcfda package available on GitHub. Assuming Gaussianity, we leverage the estimated mean and covariance functions to compute the conditional mean and variance. Subsequently, we reconstruct the underlying stochastic process using the recursive procedure outlined in equation (2.4). In our first experiment, we generate functional snippets from both the Ho-Lee model and the Ornstein-Uhlenbeck process. To obtain these snippets, we simulate the sample paths of $X_{t,i}$ over a uniform time grid $\{t_k\}_{k=0}^{K}$, where $t_k = k\delta$ and $K\delta = 1$, for each $i = 1, ..., n$, and for some constant $\delta \in (0, 1)$. The simulated values for the n processes are represented as $\{(X_{T_i,i}, X_{T_{i+\delta},i})^T\}_{i=1}^{n}$, where T_i is randomly selected from the time grid $\{t_k\}_{k=0}^{K-1}$. Since both the Ho-Lee model and the Ornstein-Uhlenbeck

process are narrow-sense linear stochastic differential equations (SDEs), we utilize exact simulation methods based on their explicit solutions, following the approach detailed in Glasserman (2004)[7].

Ho-Lee model:

The Ho-Lee model is given by the stochastic differential equation

$$\mathrm{d}X_t = g(t)\mathrm{d}t + \sigma \mathrm{d}B_t,$$

we employ a simple recursive scheme to generate values over the discrete time grid $\{t_k\}_{k=0}^{K}$,

$$X_{k+1} = X_k + \int_{t_k}^{t_{k+1}} g(s)\mathrm{d}s + \sqrt{\sigma(t_{k+1} - t_k)}.W_k, \tag{3.1}$$

where $W_k \sim \aleph(0,1)$ are independent for all k, and the initial value is set to $X_0 = x_0$.

Ornstein-Uhlenbeck process:

The Ornstein-Uhlenbeck process is described by:

$$\mathrm{d}X_t = -\theta X_t \mathrm{d}t + \sigma \mathrm{d}B_t,$$

the recursive formula takes the form:

$$X_{k+1} = e^{-\theta(t_{k+1}-t_k)}X_k + \sqrt{\frac{\sigma^2}{2\theta}(1 - e^{-\theta(t_{k+1}-t_k)})}.W_k. \tag{3.2}$$

These procedures are considered exact, meaning that the simulated values follow the same joint distribution as the corresponding continuous-time process when restricted to the simulation grid. To analyze the impact of noise, we introduce independent

errors to the generated functional snippets $\{(X_{T_i,i}, X_{T_{i+\delta},i})^T\}_{i=1}^n$.

Simulation Framework

Specifically, we simulate functional snippets $\{(Y_{i1}, Y_{i2})^T\}_{i=1}^n$contaminated with Gaussian noise:

$$Y_{i1} = X_{T_i,i} + \epsilon_{i1} \qquad Y_{i2} = X_{T_{i+\delta},i} + \epsilon_{i2},$$

with $\epsilon_{ij} \sim \aleph(0, v^2)$ being independent noise terms. The simulations consider varying sample sizes ($n = 50, 200, 1000$) and noise levels ($v = 0, 0.01, 0.1$). For each configuration, we perform $Q = 500$ simulation runs. The time interval was set to $[0, 1]$ with a uniform time step of $\delta = 0.05$.

In each simulation, we applied the recursive procedure described in equation (2.4) $M = 1000$ times using the noisy functional snippets $\{(Y_{i1}, Y_{i2})^T\}_{i=1}^n$. Starting from the initial condition $Z_0 = (0, 0)^T$. we generated $M = 1000$ estimated sample paths evaluated on the discrete time grid $\{t_k\}_{k=0}^K$. These estimated paths are denoted as:

$$\{(\hat{X}_{t_1,l}, \hat{X}_{t_K,l})^T\}_{l=1}^M.$$

For each estimated path l, we computed the corresponding true sample path $(X_{t_1,l}, X_{t_K,l})^T$ using the recursive procedure outlined in equations (3.1) or (3.2), ensuring that the same initial value and random variable W_k were used. To assess the accuracy of the approach, we quantified the estimation error for each combination of sample size and noise level using the root-mean-square error ($RMSE$).

$$RMSE = \sqrt{\frac{1}{M}\sum_{l=1}^{M}(\hat{X}_{t_K,l} - X_{t_K,l})^2}.$$

For both the Ho-Lee model and the Ornstein-Uhlenbeck process, we set the function

$g(t) = \cos(t)$, and fixed the parameters at $\theta = 1$ and $\sigma = 1$, with an initial condition of $X_0 = 0$ and M represent the number of paths or simulation runs. To estimate the conditional mean and conditional variance, we applied multiple linear regression in both cases.

Results

The simulation results, summarized in Table 3.3, which presents the mean and standard deviation (in parentheses) of the $RMSE$ across 500 simulation runs for different sample sizes and noise levels. The table compares the performance of:

- **DM:** The proposed dynamic modeling approach.
- **LW:** The covariance completion method introduced by Lin and Wang (2022) [17]

Sample Size (n)	Noise Level (v)	Ho-Lee Model (DM)	Ho-Lee Model (LW)	Ornstein -Uhlenbeck (DM)	Ornstein -Uhlenbeck (LW)
50	0	0.92 ($\pm$0.87)	1.21 ($\pm$1.71)	0.63 ($\pm$0.65)	0.71 ($\pm$1.81)
50	0.01	0.93 ($\pm$0.93)	1.08 ($\pm$0.37)	0.62 ($\pm$0.84)	0.72 ($\pm$0.23)
50	0.1	1.07 ($\pm$0.34)	1.11 ($\pm$0.45)	0.74 ($\pm$0.85)	0.74 ($\pm$0.85)
200	0	0.39 ($\pm$0.23)	0.54 ($\pm$0.38)	0.25 ($\pm$0.15)	0.27 ($\pm$0.73)
200	0.01	0.38 ($\pm$0.22)	0.91 ($\pm$0.34)	0.26 ($\pm$0.16)	0.67 ($\pm$0.17)
200	0.1	0.89 ($\pm$0.27)	0.85 ($\pm$0.27)	0.26 ($\pm$0.17)	0.67 ($\pm$0.23)
1000	0	0.17 ($\pm$0.09)	0.27 ($\pm$0.13)	0.11 ($\pm$0.06)	0.15 ($\pm$0.05)
1000	0.01	0.17 ($\pm$0.09)	0.89 ($\pm$0.24)	0.11 ($\pm$0.06)	0.17 ($\pm$0.09)
1000	0.1	0.17 ($\pm$0.09)	0.88 ($\pm$0.25)	0.11 ($\pm$0.06)	0.70 ($\pm$0.11)

Table 3.3: Mean RMSE (with Standard Deviation) Across 500Runs

The findings indicate that the proposed DM approach consistently achieves lower RMSE values as the sample size increases, suggesting improved accuracy with more data. Notably, the presence of noise has minimal impact on the performance of this method.

In contrast, the LW method exhibits significantly higher $RMSE$ values, even with a large sample size of 1000. This discrepancy may be attributed to the highly sparse nature of the simulation setup, where each process is observed within a narrow window of length 0.05. This limited observational range poses a challenge for covariance completion methods, making it difficult to accurately reconstruct the full covariance structure over the broader interval $[0, 1]$, while LW method shows some improvement when additional measurements are available, it remains less effective than the proposed DM approach.

To further demonstrate the effectiveness of the DM approach, Figure 3.2 visualizes the simulation results for the Ornstein-Uhlenbeck process, considering a sample size of $n = 200$ and a noise level of $v = 0.1$. The figure includes 100 estimated sample paths alongside the corresponding true sample paths. The strong alignment between the estimated and true paths highlights the ability of the proposed method to accurately capture the underlying stochastic dynamics, even when working with sparse data.

The following figures illustrate the simulated paths of the Ho-Lee and Ornstein-Uhlenbeck models:

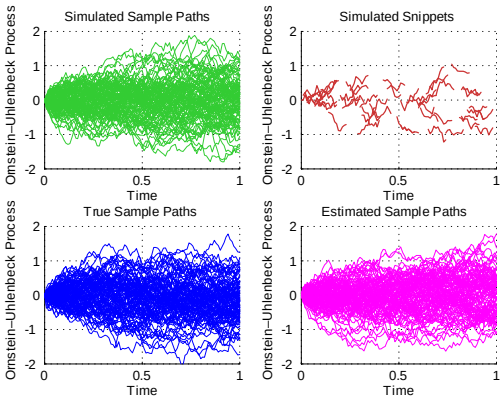

Figure 3.2: Simulated Paths of the Ornstein-Uhlenbeck Model

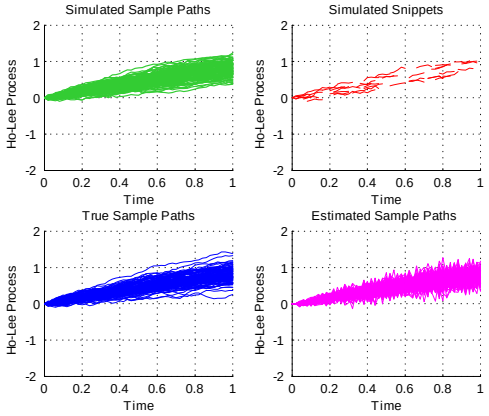

Figure 3.3: Simulated Paths of the Ho-Lee Model

This chapter demonstrated the effectiveness of the stochastic differential equation (SDE) approach in modeling sparse longitudinal data, particularly in the Nepal Growth Study. The results showed that the model effectively handled missing data and accurately predicted growth patterns. Simulation studies confirmed that the proposed method outperformed traditional approaches by reducing errors. This work paves the way for future applications in financial modeling and climate data analysis.

Conclusion

This thesis presented a robust framework for modeling sparse longitudinal data using stochastic differential equations (SDEs), addressing the research problem by demonstrating their capacity to reconstruct underlying stochastic processes and improve estimation accuracy without relying on strong assumptions.

Three key results emerged:

1. **Theoretical Foundation:** Established conditions for the existence and uniqueness of SDE solutions, proving their applicability in dynamic systems through models such as the Ornstein-Uhlenbeck process.*

2. **Methodological Innovation:** Introduced novel simulation-based estimation techniques that outperformed traditional methods (e.g., FPCA) in analyzing sparse and irregularly sampled data.

3. **Empirical Validation:** Confirmed the framework's effectiveness through a case study on child growth patterns in Nepal and simulations using Ho-Lee and Ornstein-Uhlenbeck models, demonstrating accurate dynamic distribution reconstruction despite missing data.

By developing this computational framework, the study advances prediction accuracy and dynamic phenomena analysis across disciplines. Future research could extend

this work through integration of nonlinear stochastic processes, non-Gaussian distributions, and deep learning techniques, broadening its applications in medicine, economics, and social sciences..

Annex A: Lemma and Proof

Lemma 3.2.1

$$\mathbb{E}(X_s|X_t) = \mu(s) + \sum(s,t)\sum^{-1}(t,t)\{X_t - \mu(t)\}, \tag{3.3}$$

$$Var(X_s|X_t) = \sum(s,s) - \sum(s,t)\sum^{-1}(t,t)\sum(t,s). \tag{3.4}$$

Proof. For all $s \geq t$, based on the fact that the random variable $\mathbb{E}(X_s|X_t)$ is Gaussian, we aim to find the best linear estimate of $\mathbb{E}(X_s|X_t)$, meaning that we seek a linear function of the form:

$$\mathbb{E}(X_s|X_t) = a + bX_t.$$

Then,

$$\mathbb{E}(X_s) = a + b\mathbb{E}(X_t),$$

i.e,

$$\mu(s) = a + b\mu(t).$$

Then,

$$a = \mu(s) - b\mu(t).$$

This leads to,

$$\mathbb{E}(X_s|X_t) = \mu(s) + b(X_t - \mu(t)).$$

Hence,

$$\mathbb{E}(X_s|X_t) - \mu(s) = b(X_t - \mu(t)).$$

So,

$$\mathbb{E}[(\mathbb{E}(X_s|X_t) - \mu(s))(\mathbb{E}(X_t|X_t) - \mu(t))] = \mathbb{E}[b(X_t - \mu(t))^2].$$

It implies that

$$\sum(s,t) = b\sum(t,t).$$

Consequently,

$$b = \frac{\sum(s,t)}{\sum(t,t)} = \sum(s,t)\sum^{-1}(t,t).$$

Thus,

$$\mathbb{E}(X_s|X_t) = \mu(s) + \sum(s,t)\sum^{-1}(t,t)(X_t - \mu(t)). \tag{3.5}$$

Since $\mu(s)$, $\sum(s,t)$, $\sum^{-1}(t,t)$ are all measurable functions, and X_t is also measurable, the expression (3.5) is entirely $\sigma(X_t)-$measurable. Therefore, when taking the conditional expectation $\mathbb{E}(X_s|X_t)$, it remains unchanged, leading to:

$$\mathbb{E}(X_s|X_t) = \mu(s) + \sum(s,t)\sum^{-1}(t,t)(X_t - \mu(t)).$$

Now, I will prove (3.4) based on (3.3).We want to find:

$$\mathbb{E}[Var(X_s|X_t)].$$

We know that:

$$Var(X_s) = \mathbb{E}[Var(X_s|X_t)] + Var(\mathbb{E}(X_s|X_t)).$$

Thus, by rearranging the equation, we obtain:

$$\mathbb{E}[Var(X_s|X_t)] = Var(X_s) - Var(\mathbb{E}(X_s|X_t)). \tag{3.6}$$

Now, from equation (3.3), we have:

$$\mathbb{E}(X_s|X_t) = \mathbb{E}[X_s] + \frac{cov(X_s, X_t)}{var(X_t)}(X_t - \mathbb{E}[X_t]).$$

Taking the variance on both sides, we get:

$$Var(\mathbb{E}(X_s|X_t)) = (\frac{cov(X_s, X_t)}{var(X_t)})^2 Var(X_t),$$

which simplifies to:

$$Var(\mathbb{E}(X_s|X_t)) = \frac{cov^2(X_s, X_t)}{var(X_t)}.$$

Finally, substituting this result into equation (3.6), we conclude that:

$$\mathbb{E}[Var(X_s|X_t)] = Var(X_s) - \frac{cov^2(X_s, X_t)}{var(X_t)}.$$

Then,

$$Var(X_s|X_t) = Var(X_s) - \frac{cov^2(X_s, X_t)}{var(X_t)}.$$

i.e,

$$Var(X_s|X_t) = \sum(s,s) - \sum(s,t)\sum^{-1}(t,t)\sum(t,s).$$

■

Bibliography

[1] Castro, P. E., Lawton, W. H., & Sylvestre, E. A. (1986). Principal modes of variation for processes with continuous sample curves. Technometrics, 28(4), 329–337.

[2] Chen, K., & Lei, J. (2015). Localized functional principal component analysis. Journal of the American Statistical Association, 110(512), 1266–1275.

[3] Chen, K., & Müller, H.-G. (2012). Conditional quantile analysis when covariates are functions, with application to growth data. Journal of the Royal Statistical Society: Series B, 74(1), 67–89.

[4] Dawson, M., & Müller, H.-G. (2018). Dynamic modeling of conditional quantile trajectories, with application to longitudinal snippet data. Journal of the American Statistical Association, 113(523), 1612–1624.

[5] Evans, L. C. (2013). An introduction to stochastic differential equations. American Mathematical Society.

[6] Fan, J., & Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. Biometrika, 85, 645–660.

[7] Glasserman, P. (2004). Monte Carlo methods in financial engineering (Vol. 53). Springer, New York, NY.

[8] Hall, P., & Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), 109–126.

[9] Hall, P., & Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. Annals of Statistics, 35(1), 70–91.

[10] Hsing, T., & Eubank, R. (2015). Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons.

[11] Kloeden, P. E., & Platen, E. (1999). Numerical solution of stochastic differential equations (Vol. 23). Springer, Berlin, Heidelberg.

[12] Kleffe, J. (1973). Principal components of random variables with values in a separable Hilbert space. Mathematische Operationsforschung und Statistik, 4(4), 391–406.

[13] Labed, B. (2024). Équations différentielles stochastiques [Unpublished lecture notes, Master 2]. Université Mohamed Khider, Biskra.

[14] Le Gall, J.-F. (2016). Mouvement brownien, martingales et calcul stochastique. Springer. (Mathématiques et Applications, Vol. 71).

[15] Lin, Z., & Wang, J.-L. (2022). Mean and covariance estimation for functional snippets. Journal of the American Statistical Association, 117(538), 348–360.

[16] Ramsay, J. O., & Silverman, B. W. (2005). Functional data analysis (2nd ed.). Springer.

[17] Villani, C. (2009). Optimal transport: Old and new (Vol. 338). Springer, Berlin, Heidelberg.

[18] Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2016). Functional data analysis. Annual Review of Statistics and Its Application, 3, 257–295.

[19] West Jr, K. P., LeClerq, S. C., Shrestha, S. R., Wu, L. S.-F., Pradhan, E. K., Khatry, S. K., Katz, J., Adhikari, R., & Sommer, A. (1997). Effects of vitamin A on growth of vitamin A-deficient children: Field studies in Nepal. Journal of Nutrition, 127(9), 1957–1965.

[20] Zhou, H., Wei, D., & Yao, F. (2022). Theory of functional principal components analysis for discretely observed data. arXiv preprint arXiv:2209.08768.

[21] Zhou, Y., & Müller, H. G. (2024). Dynamic modelling of sparse longitudinal data and functional snippets with stochastic differential equations. Journal of the Royal Statistical Society Series B: Statistical Methodology, qkae116.

ملخص :

تستخدم النماذج الرياضية البيانات الطولية لتحليل تطور الظواهر، لكن البيانات المبعثرة تشكل تحديًا بسبب الملاحظات غير المنتظمة. تقترح هذه الأطروحة نموذجًا ديناميكيًا يعتمد على المعادلات التفاضلية العشوائية لتحليل هذه البيانات بطريقة غير بارامترية. يتم التحقق من فعالية النموذج من خلال تطبيقه على بيانات نمو الأطفال في نيبال ومن خلال المحاكاة باستخدام نماذج هو-لي وأورنشتاين-أوهلينبيك، مما يوضح قدرته على إعادة بناء التوزيعات الديناميكية على الرغم من البيانات غير المكتملة

الكلمات المفتاحية: البيانات الطولية المتفرقة، المعادلات التفاضلية العشوائية، التقدير غير المعلمي، المحاكاة .

Abstract:

Mathematical models utilize longitudinal data to analyze the evolution of phenomena, but sparse data presents a challenge due to irregular observations. This thesis proposes a dynamic model based on stochastic differential equations to analyze this data in a non-parametric manner. The model's effectiveness is validated through its application to child growth data in Nepal and simulations using the Ho-Lee and Ornstein-Uhlenbeck models, demonstrating its ability to reconstruct dynamic distributions despite incomplete data.

Key words: **Sparse longitudinal data, Stochastic differential equations, Nonparametric estimation, Simulation.**

Résumé:

Les modèles mathématiques utilisent des données longitudinales pour analyser l'évolution des phénomènes, mais les données éparses posent un défi en raison des observations irrégulières. Ce mémoire propose un modèle dynamique basé sur les équations différentielles stochastiques pour analyser ces données de manière non paramétrique. L'efficacité du modèle est validée par son application aux données de croissance des enfants au Népal et par des simulations utilisant les modèles Ho-Lee et Ornstein-Uhlenbeck, démontrant sa capacité à reconstruire des distributions dynamiques malgré des données incomplètes.

Mots clés : **Données longitudinales éparses, Équations différentielles stochastiques, Estimation non paramétrique, Simulation.**